一般口演
# 一般口演28
# 機械学習・アルゴリズム・解析モデル
2017年11月23日(木) 12:45 〜 14:15 E会場 (10F 会議室1003)

[4-E-2-OP28-3] 機械学習アルゴリズムによる効率的なリアルワールドデータ解析を実現する手法の提案

早川 龍雄[1], 三木 瑞穂[1], 中津井 雅彦[2], 内野 詠一郎[2], 種石 慶[3], 奥野 恭史[2] （1.富士通株式会社 ヘルスケアシステム事業本部 イノベーション推進事業部, 2.京都大学大学院医学研究科, 3.理化学研究所）

【目的】
病院情報システムにて蓄積され続ける医療ビッグデータ（リアルワールドデータ）は、多種多様でかつ構造は複雑で、多くの欠測値・バイアス・外れ値を含んでいる。従来の解析手法を利用しても有意性のある結果を出力することは困難であり、データクレンジング技術が必須である。データが整理された後も機械学習においては最適なアルゴリズムの選択、及びハイパーパラメータの探索など多様な技術を駆使した処理の実現が必須である。一般的には個々の研究テーマに応じて試行錯誤的にそれらの処理を実施・評価する為、多くの時間・コストを費やす。データ解析に必要な処理を予め部品化し、内部的にデータ構造を解析することで、最適な部品を組み合わせて、解析モデルを作成する為のフローを推薦する仕組みを提案する。
【方法】
上記提案手法について有効性を立証する為、データ解析システムのプロトタイプを開発し、過去に実施された研究テーマにおける解析プロセスと比較してどれくらい作業効率化されたかという点と正しく結果が出力されたかという点について評価した。病名・血液検査・副作用の時系列データをインプットとして、データ構造からクレンジングの変換定義を自動的に作成・推薦を行い、その後に最適な解析アルゴリズムとハイパーパラメータを抽出した上でモデルを作成し、評価用の PLOT図を作成する一連の解析フローについて検証した。
【結果】
データクレンジングの実施から解析結果出力までの一連の処理を試行錯誤することにかかる時間を今回のデータ解析システムでは数日単位から数時間単位に削減することが出来る。今まで解析が困難な欠損値が多い組合せデータ群の利用価値を高め、統計的な傾向分析でなく、ビッグデータが持つ実際の値の組合せからの類似性や特徴性の探索とモデル化が可能となった。今後、出力されたモデルの組合せにより医療現場で利用できるレコメンドシステムへの応用が期待される。

機械学習アルゴリズムによる
効率的なリアルワールドデータ解析を実現する手法の提案

早川　龍雄[*1]、三木　瑞穂[*1]、
中津井　雅彦[*2]、内野　詠一郎[*2]、種石　慶[*3]、奥野　恭史[*2]
*1 富士通株式会社　ヘルスケアシステム事業本部　イノベーション推進事業部、
*2 京都大学大学院医学研究科、*3 理化学研究所

Proposal of technique for achieving efficient, "Real World Data" analysis by machine-learning algorithm

Hayakawa Tatsuo[*1], Miki Mizuho[*1],
Nakatsui Masahiko[*2], Uchino Eiichiro[*2], Taneishi Kei[*3,] Okuno Yasushi[*2]
*1 FUJITSU Ltd. healthcare system business headquarters innovation promotion division,
*2 Graduate School of Medicine, Kyoto University, *3 RIKEN

The medical treatment big data (Real World Data) that keeps being accumulated with the hospital information system contains a lot of missing values, biases, and unexpected values. It is difficult to output the meaningful result with the significance even if a past method for analyzing is used. Processing to make good use of various technologies (data preparation and searching for hyper parameter the best algorithm, for using machine-learning, etc.) is important. Generally, to execute and to evaluate it while doing the trial and error of analytical processing to individual topics of research, a lot of time and the cost are spent. We developed the prototype of the data analysis system to make analytical flow efficiently, how much made a efficiency compared with past topics of research was verified. A flow of processing from the execution of the data preparation to the analytical result output and the time that hung to the trial and error, was able to be reduced greatly in this data analysis system. It became possible to analyze and modeling in raising the utility value of the data including a lot of difficult missing values. The application to recommend-system that can be used by combining the output model on the medical treatment site will be expected in the future.

Keywords: Big Data, Real World Data, Machine-Learning, Data analysis system,

1. はじめに

国の健康医療戦略である「健康長寿社会」の実現に向け、先制医療・個別化医療に有用なゲノム情報を含めた医療データの解析は、全国の大学機関や国立研究機関（ナショナルセンター）を中心として研究が進められている。医療・創薬の発展だけでなく、新産業の創出という側面においても産学官民が連携して医療データの解析に取り組んでいくことが期待されている[1]。

医療データは過去からの蓄積分を含めて膨大な「医療ビッグデータ＝リアルワールドデータ」となっているが、これらのデータは医療現場での多種多様な情報の集合体であることから、単純に集積し整理・解析を実施することは困難である。健康長寿社会の実現に向けた研究を進める上で、有効な解析が可能となる情報基盤の整備は課題となっている。データ解析の作法・手順は研究者が手探りで試行錯誤的に実施しているのが現状であり、研究現場では効率的な仕組み作りへの対応が求められている[2]。

我々は、2004 年からの 10 年間の約 5200 人の匿名化されたがん患者の実臨床データを用いて、医薬品の薬効・副作用のパターンの時系列解析を実施している。その中で、効率的な医療ビッグデータ解析における課題の解決と研究作業の効率化を目指し、研究現場で簡易に利用できるデータ解析システムのプロトタイプを開発し、評価を行っている。このシステムにおいて重要な課題解決テーマであるデータクレンジングとリアルワールドデータの解析手法について、プロトタイプとして実装内容を踏まえて考察する。

2. データプレパレーションの重要性

多種多様でかつ複雑な構造を持つ医療データを単純に集積し解析を行っても有意性のある結果を出力することは困難である。解析においては、データのバイアスや欠測値の取扱いを考慮したデータプレパレーションが必須である。特にデータクレンジングに関するプロセスは、解析作業全体の時間の約 8 割を占める作業と言われており、効率的なデータ解析を実施する上で重要である[3]。実際に著者がある研究テーマにおいて実施した作業の割合は図1のようになった。

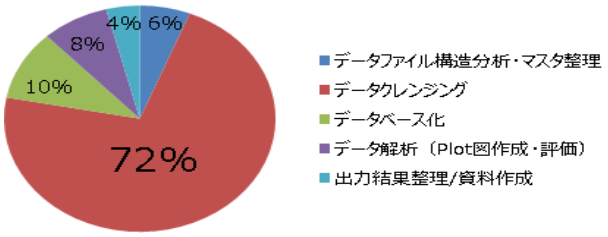

図 1：データ解析時間の内訳

がん患者の各種実臨床データの解析を実施しながら、課題の抽出と効率的な解析プロセスの検討を実施してきた。当初は一般的な統計解析の手法である程度のデータ内の相関や特徴が見られると考えていたが、実際にはデータ内に多くの課題があり、正確な解析結果を阻害する要因となっていた。

血液検査結果の場合には、毎回同じ検査項目のデータを取得しているわけではなく、かつ時系列として一定のインターバルでデータが存在していないという特殊な欠測値となって

いるため、データの補間処理が必要となる。その時に、単純であれば 0 値や平均値などの補間アルゴリズムを利用するのだが、0 値による補間はデータとしてバイアスとなるため不向きであり、また複数の患者の時系列データに平均値などはデータの意味が変わってしまう為、適用できない。対応として患者別にデータグルーピングを行い、その時系列データの並びと前後の値の差から線形による予測値を設定するスプライン補間などを適用することを実施した。患者によっては、特定の期間内で病状の再発などにより長期間の治療空白期間などが存在するケースもあり、同じ患者でも治療期間によってデータの集合を分割するというデータフレームの構築作業も必要になる。

データが存在していても、検査データは複雑な数値表記になっており、ほとんどの場合に正規表現としての構造になっておらず、外れ値により正確な特徴抽出を阻害することが多い。統計処理で一般的に利用される四分位数による外れ値除去を正規表現になっていないデータに対して実施すると、図2のように過剰にデータが削除され、データの特徴性が丸められることで正しい結果を取得できなくなるケースもある。このようにデータプレパレーション処理は過不足なく根拠に基づいて実施されることが必要であり、システム化を行う上での重要である。

元データの統計量

項目	有効データ数	最小値	第1四分位	中央値	第3四分位	最大値
A	382110	0.0	NaN	NaN	NaN	27.7
B	382108	0.7	NaN	NaN	NaN	10.3

欠測値補間後の統計量

項目	有効データ数	最小値	第1四分位	中央値	第3四分位	最大値
A	427121	0.0	9.8	11.4	12.8	27.7
B	427121	0.7	29.3	30.8	32.4	300.0

外れ値除去後の統計量

項目	有効データ数	最小値	第1四分位	中央値	第3四分位	最大値
A	21137	5.9	11.1	12.2	13.3	17.1
B	21137	24.8	29.5	30.8	32.0	36.9

図 2:データ補間および外れ値除去実施前後の統計量

対応として、通常ではデータ件数を見ながら最適なサンプリング数を確保する閾値やカットオフ値を探索することが多く、試行錯誤的に実施することに時間がかかる。今回検討した手法では、あらかじめ定義した閾値やデータ補間アルゴリズムの組合せを仮実行し、有効データ数の割合から最適な外れ値除去することを実施した。結果はヒストグラムや散布図で確認することができ、スクリプトを作りこまなくても対応が可能になった。

ある程度データプレパレーションはシステムとして汎用的に対応することが可能になるが、実際は医療固有のデータ概念の知識(ドメインナレッジ)が必要であり、全て自動化することはリスクにもなりえる。システムにおいては自動的な処理を考慮しながらも、パラメータにより研究者の主観性を反映し、作業の履歴が参照・復元できる仕組みが重要であると考える。

3. 変数設定とデータセット作成

データプレパレーションを実施したデータに対して、様々なデータ解析を実施するが、今回の検討では「クラス分類」「回帰分析」「クラスタリング」「時系列予測」のモデル構築を主軸とした解析フロー自動作成を行う方針とした。

医療データはあるデータを判断するラベル(目的変数)に影響を与える多くの項目(説明変数)が存在しており、どれが必要な項目かを判断することは困難である。取得した全ての項目を説明変数とすることもあるが、説明変数間の結合度に左右され、相関性だけが強調されることにより、目的変数に対する特徴を探索できない場合がある。また、医師の知見や論文の記載内容に準拠して関連性が高い説明変数を設定する方法も、隠れた因果関係の探索という点では不向きであると考える。

説明変数の探索として、ベイジアンネットワークによる解析を行い、出力された因果関係をグラフ化し、相関分析を実施した結果と合わせて、目的変数を主軸とした説明変数の抽出を実施する方針とした。この仕組みは、比較的単純な処理により代表的な項目を抽出することが出来る上、それぞれの項目の視点を図3のように変えて表示することで、目的変数と説明変数の関連を簡単に把握することが出来る。さらに、目的変数をドリルダウンし、図4のように有向グラフ化することで影響する方向と相関値の強弱を確認することも可能である。

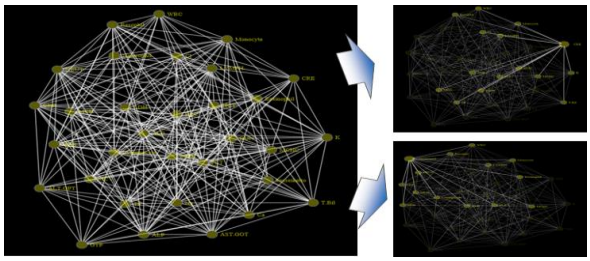

図 3:グラフィカルモデルによる説明変数探索イメージ

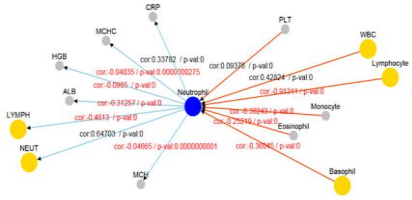

図 4:有向グラフによる影響する方向と相関性表示イメージ

データ解析対象を抽出後、解析用のデータセットを作成するが、時系列のデータかどうかによりデータセットの作成方法が異なる。時系列データの場合は、特定の時点(エンドポイント)を事前に定義する必要があり、その時点前後で一定の時間軸を自動的に作成し、データ項目としている。エンドポイントをイベントとする時間の遡りにより、起因するデータの特徴を抽出し、モデル化することで予後予測に活用が出来ると考える。時系列データでない場合は、目的変数と説明変数をマトリクス上にデータセットを作成する。ただ、複数の項目において、データの変量に大きな違いがある場合は、変量が大きい項目に影響され、正しく解析が出来ないケースがある。その為、一度仮に階層クラスタリングとして図5のように出力し、データ標準化が必要かどうかを判定し、必要であれば標準化処理を行うようにしている。

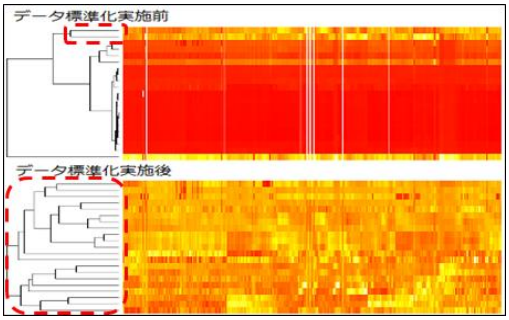

図 5:データ標準化実施前後比較

4. アルゴリズム・ハイパーパラメータ探索

機械学習において、処理するデータ量・得られる精度・実行時間など多様な制約の中で、アルゴリズムやハイパーパラメータの膨大な組合せの中から最適なものを選択することは困難である。限られた時間制約下で巨大データに対し複数の学習アルゴリズムの候補の中から最も精度の高いものを高速に自動探索する仕組みとして、サブサンプリングを利用した富士通で提唱している wizz を最適な選択肢を自動的に選ぶプラットフォームとして利用している[4]。

従来にもアルゴリズムやハイパーパラメータを探索する技術は存在していたが、今回着目した点は、医療データとして大量のデータ件数で、かつ説明変数としての項目数が多い場合に以下に短時間に処理できるかという点である。Wizz ではプログレッシブサンプリング法を複数のアルゴリズムに跨って並行実行しており、全訓練データによる精度を最小のサンプルサイズで得ることができ、ビッグデータでも高速に処理できる点がメリットと考える。

この技術を利用して、オープンデータによる高血圧患者かどうかのラベルを利用したクラス分類の擬似テーマで検証を行った。データ件数は 2 万件で説明変数を約 50 に設定し、wizz によるアルゴリズムとハイパーパラメータ探索、モデルの構築を実施したところ、一連の作業がほぼ 10 分弱程度で完了し、ROC 曲線における AUC も 0.95 という結果を得ることができた。従来の試行錯誤的な検証に比べ、作業効率化が実現できたと考える。

5. データ解析プラットフォームの検証

データプレパレーションからモデル作成までの一連の処理をさらに効率化するため、GUI で解析フローを作成するデータ解析用システムのプロトタイプを開発した。インタフェースは図 6 のような WEB アプリケーションとしており、今後の検討でクラウドでの利用を想定している。利用者は、アカデミアの研究者を想定しているが、Python や R のスクリプトの構築経験がない利用者でもパラメータで設定した項目を元に自動的にスクリプトを作成する仕様としており、統計処理に詳しくなくても利用できるような仕組みとして構築した。

図 6:データ解析用システムのプロトタイプ(イメージ)

このプロトタイプを利用して、それまでに実施した研究テーマを解析し、どれくらい効率化できるかを検証した。大きく異なる点は、Python や R のスクリプトをシステムでは記載しなくてもよいという点であり、ライブラリのパラメータを画面から設定できるので、試行錯誤的に処理を行う場合でも簡単に設定値を変更可能となる。試行錯誤におけるパラメータ設定値と自動作成されたスクリプト、出力した結果とログは自動的に紐付け保存されるので、解析フローのトレーサビリティ管理も意識せずに保存できる。ソースとデータの運用管理を自動管理できることで、その分の作業の手間を大幅に改善できる。

解析としては、血液検査結果データにおける好中球減少と治療イベントの関連を研究テーマとして検証を行った。検証データとしては、好中球減少の発生有無をカテゴリ変数化し、発生前 30 日前の治療イベントと他の血液検査項目を利用した。以前の解析では、データの組合せやアルゴリズムの比較などで、毎回かなりの時間を費やしており、探索に約1ヶ月かかっていた。今回のシステムでは、その点がほぼ自動化されていることで、数時間の処理で対応が可能となった。処理時間を短縮することが、リアルワールドデータによる解析を実現可能とする為に重要であると考える。

6. まとめ

本論文では、機械学習技術の組合せにより、リアルワールドデータの効率的な解析の実施とモデルの作成について述べた。

今回構築したプロトタイプは、今後クラウド上で利用するサービスとして提供を目指しており、このデータ解析基盤で作成したモデルをデプロイし、アンサンブル学習することで AI としてのインタフェースを提供することを目指している。その為には常にデータが蓄積・再構成されるリアルワールドデータを常に監視し、自動的にモデル作成と評価を行う仕組みにして行くことが重要と考える。

リアルワールドデータの解析は、クラスタリングや機械学習を利用した手法が主流になると考える。例えば、複数の波形情報から変化状態の累積度が強い予兆波形パターンを探索する検査値変動のシミュレーション予測モデルの活用により、病気の悪化防止アラームを臨床現場で利用する検討も研究テーマの一つとなっている。様々な研究により生み出されるモデルが将来的に電子カルテ上で医師に対して効果的治療法としてサジェッションされることで、医療品質と患者の QOL (Quality Of Life)の向上が期待できると考える。

参考文献

1) 医療ビッグデータ・コンソーシアム 21 世紀医療フォーラム,
政策提言 2015, 2015 年 12 月 14 日, p.3
http://www.medical-bigdata.jp/pdf/teigen151214.pdf
2) 京都大学大学院医学研究科 中山 健夫,
「「医療ビッグデータ時代」の幕開け」, 医学書院,
http://www.igaku-shoin.co.jp/paperDetail.do?id=PA03107_02
3) 倉橋一成,「図解&事例で学ぶビジネス統計教科書」, 株式会社マイナビ, 2015 年 7 月 31 日,
初版第1版, ISBN978-4-8399-5442-0, p.90
4) 富士通研究所,小林健一,浦晃,上田晴康,
「サブサンプリングを用いた高時間高率な学習アルゴリズム選択」、電子情報通信学会 信学技報 IEICE Technical Report、
[PRMU2015-73, IBISML2015-33(2015-09)]